МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
имени М. В. Ломоносова

Механико - Математический факультет
Кафедра теории вероятностей

КУРСОВАЯ РАБОТА

«Восстановление пропущенных значений в эпидемиологических исследованиях. Методы и их реализация»

Выполнила студентка группы 408
кафедры теории вероятностей
Меркушина А. В.

Научный руководитель:
доктор физико-математических наук, профессор
Яровая Е. Б.

Москва, 2021

Содержание

Введение

Обработка и анализ данных, имеющих пропущенные значения, являются актуальным вопросом прикладной математической статистики. Например, выборки, собранные в больших эпидемиологических исследованиях, зачастую содержат пропущенные значения [2]. При предварительной обработке данных исследователь вынужден либо восстановить пропуск в наблюдениях, либо удалить участника, имеющего потерянное значение. Замена средним, медианой или константой может оказаться некорректной [4]. Также удаление всех наблюдений, связанных с утерянным, часто недопустимо из-за того, что объём выборки может значительно сократиться. По этим причинам в последнее десятилетие развились подходы, задача которых — восстановить совместное распределение переменных выборки и подобрать для пропущенного значения наиболее правдоподобное заполнение [4].

Цель работы — изучить теоретическую основу, необходимую для применения метода восстановления пропущенных данных, а также сравнить эффективность метода заполнения гиббсовским семплированием с другими методами. Для этого в главе 1 сделан обзор типов пропущенных данных, а также методов симуляции пропусков. В главе 2 приведены необходимые сведения из теории однородных марковских цепей с дискретным числом состояний и марковских процессов с непрерывным множеством состояний, сформулирована эргодическая теорема. В главе 3 приведен алгоритм гиббсовского семплирования [8] — один из методов Монте-Карло по схеме марковской цепи. Этот метод является одним из теоретических обоснований алгоритма заполнения пропущенных данных. Кроме того, в главе три нами предложен пример применения Гиббсовского семплирования к задаче моделирования выборки из многомерной случайной величины, которую нельзя решить с помощью функциональных преобразований равномерной случайной величины. В главе 4 описывается сам метод множественного вменения MICE и изучается эффективность данного метода.

1 Пропущенные данные

1.1 Типы пропущенных данных

Пусть X — матрица данных x_{ij} с пропущенными значениями,
M — матрица, хранящая позиции пропущенных данных в X :

$$M_{ij} = \begin{cases} 0, \text{ если } x_{ij} \text{ — пропущенное значение,} \\ 1, \text{ в противном случае.} \end{cases} \tag{1}$$

X_{obs} — наблюдаемые данные,
X_{mis} — пропущенные данные,
θ — параметр модели данных с отсутствующими значениями [1].

Тогда *модель данных с отсутствующими значениями* будет описываться выражением:

$$P\left(M_{ij} = 0 \mid X_{iobs}, X_{imis}, \theta\right) \tag{2}$$

MCAR (missing completely at random): отсутствие значений бессистемно, не зависит от какой-либо другой переменной в наборе данных.

$$P\left(M_{ij} = 0 \mid X_{iobs}, X_{imis}, \theta\right) = P\left(M_{ij} = 0 \mid \theta\right) \tag{3}$$

Примеры:

- данные о респондентах отсутствуют вследствие утери их анкет (Рис. 1.(2));
- измерения артериального давления могут отсутствовать из-за поломки тонометра.

MAR (missing at random): причина потери данных связана с наблюдаемыми значениями.

$$P\left(M_{ij} = 0 \mid X_{iobs}, X_{imis}, \theta\right) = P\left(M_{ij} = 0 \mid X_{iobs}, \theta\right) \tag{4}$$

Примеры:

- вероятность пропуска покахателя IQ связана с возрастом: значения отсутствуют у людей старше 34 лет (Рис. 1.(3));
- пропущенные измерения артериального давления могут быть ниже, чем измеренное артериальное давление: у молодых людей больше шансов иметь отсутствующие показатели.

MNAR (missing not at random): отсутствие значений зависит как от наблюдаемых, так и от пропущенных данных.

$$P\left(M_{ij} = 0 \mid X_{iobs}, X_{imis}, \theta\right) = P\left(M_{ij} = 0 \mid \theta\right) \tag{5}$$

Примеры:

- показатели IQ отсутствуют у людей с низкими значениями (Рис. 1.(4));

- люди с депрессией с большей вероятностью будут пропускать вопросы о ментальном состоянии.

	Age	IQ score
0	22	90
1	23	71
2	23	87
3	25	72
4	26	75
5	29	93
6	30	101
7	34	99
8	35	118
9	39	112
10	40	122
11	48	130
12	51	105
13	52	122
14	54	117

	Age	IQ score
0	22	90
1	23	71
2	23	nan
3	25	nan
4	26	75
5	29	93
6	30	nan
7	34	99
8	35	118
9	39	112
10	40	nan
11	48	130
12	51	105
13	52	nan
14	54	117

	Age	IQ score
0	22	90
1	23	71
2	23	87
3	25	72
4	26	75
5	29	93
6	30	101
7	34	99
8	35	nan
9	39	nan
10	40	nan
11	48	nan
12	51	nan
13	52	nan
14	54	nan

	Age	IQ score
0	22	90
1	23	nan
2	23	87
3	25	nan
4	26	nan
5	29	93
6	30	101
7	34	99
8	35	118
9	39	112
10	40	122
11	48	130
12	51	105
13	52	122
14	54	117

Рис. 1. (1) full (2) MCAR (3) MAR (4) MNAR

1.2 Симуляция однофакторной потери данных

Прежде чем перейти к рассмотрению методов заполнения, необходимо понять, каким образом будет оцениваться их эффективность. Для оценки метода заполнения необходимо сравнить новые (заполненные) значения с исходными (настоящими). Для этого будем брать полный датасет и симулировать в нем отсутствующие значения по трем типам пропущенных данных: MCAR, MAR, MNAR.

Пусть есть датасет с параметрами X_1, X_2, X_3, X_4, X_5. Рассмотрим вероятности потерять значения по параметру X_1 по трем типам:

1. Потеря данных по типу MCAR случайна, поэтому потерянные значения будем симулировать с помощью распределения Бернулли с вероятностью 0.5;

2. Вероятность потерять значения по типу MAR зависит только от наблюдаемых данных, то есть от параметров X_2, X_3, X_4, X_5. Для прогнозирования вероятности наступления пропуска используем логистическую регрессию:

$$f(z) = \frac{1}{1 + e^{-z}}, \qquad z = 0 \cdot X_1 + \theta_2 \cdot X_2 + \theta_3 \cdot X_3 + \theta_4 \cdot X_4 + \theta_5 . \cdot X_5 \tag{6}$$

   Коэффициенты регрессии $\theta_2, \theta_3, \theta_4, \theta_5$ положим равными $-0.08, 0.4, 0.2, -0.45$;

3. Вероятность потерять значения по типу MNAR зависит и от наблюдаемых, и от пропущенных данных, то есть от всех параметров X_1, X_2, X_3, X_4, X_5. Для прогнозирования вероятности наступления пропуска используем логистическую регрессию:

$$f(z) = \frac{1}{1+e^{-z}},$$
$$z = \theta_1 \cdot X_1 + \theta_2 \cdot X_2 + \theta_3 \cdot X_3 + \theta_4 \cdot X_4 + \theta_5 \cdot X_5 \tag{7}$$

Коэффициенты регрессии $\theta_1, \theta_2, \theta_3, \theta_4, \theta_5$ положим равными $-1, -0.08, 0.4, 0.2, -0.45$.

Тогда полные данные (оранжевые точки) и потерянные данные (синие точки):

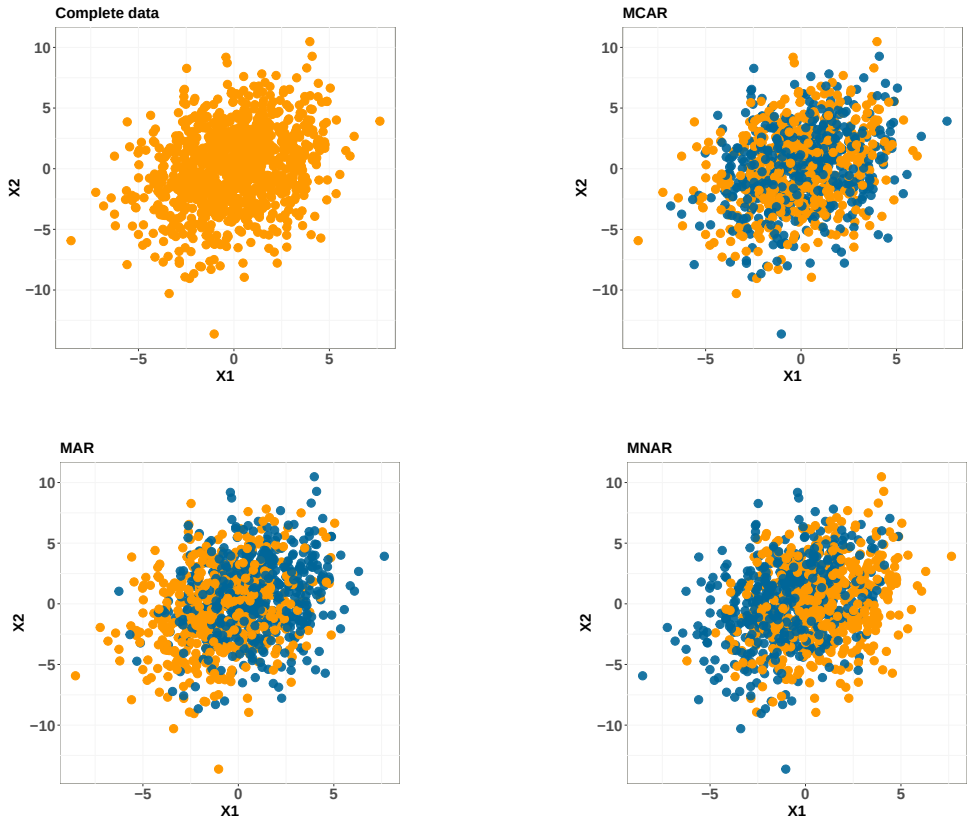


Рис. 2. Зависимость параметров данных X_2 от X_1 при разных типах потери данных.

Посмотрим, как потеря значений влияет на устройство данных.

Для этого сгенерируем 1000 выборок полных данных и по 1000 выборок данных с пропущенными значениями по MCAR, MAR, MNAR. Для каждого датасета посчитаем корреляцию между параметрами X_1 и X_2, после чего найдем среднее значение корреляции.

В итоге получим следующие значения корреляций для полных данных и для тех же данных со сгенерированными пропусками, из которых видно, что потеря значений может как ослабить связь между параметрами (Рис.3 MAR), так и усилить ее (Рис.3 MNAR):

cor_fulldate_X1X2	cor_MCAR_X1X2	cor_MAR_X1X2	cor_MNAR_X1X2
0.2466219	0.2457660	0.1903382	0.2717526

Рис. 3. Среднее значение корреляции между параметрами X_1 и X_2.

cor_fulldate_X1X2	cor_MCAR_X1X2	cor_MAR_X1X2	cor_MNAR_X1X2
0.02964046	0.04189263	0.04340688	0.04289280

Рис. 4. Стандартное отклонение корреляции между параметрами X_1 и X_2.

1.3 Симуляция многофакторной потери данных. Устройство метода ampute

Однако на практике нас интересует не однофакторная потеря данных, а реалистичная - многофакторная. Мы будем симулировать ее с помощью функции ampute [3].

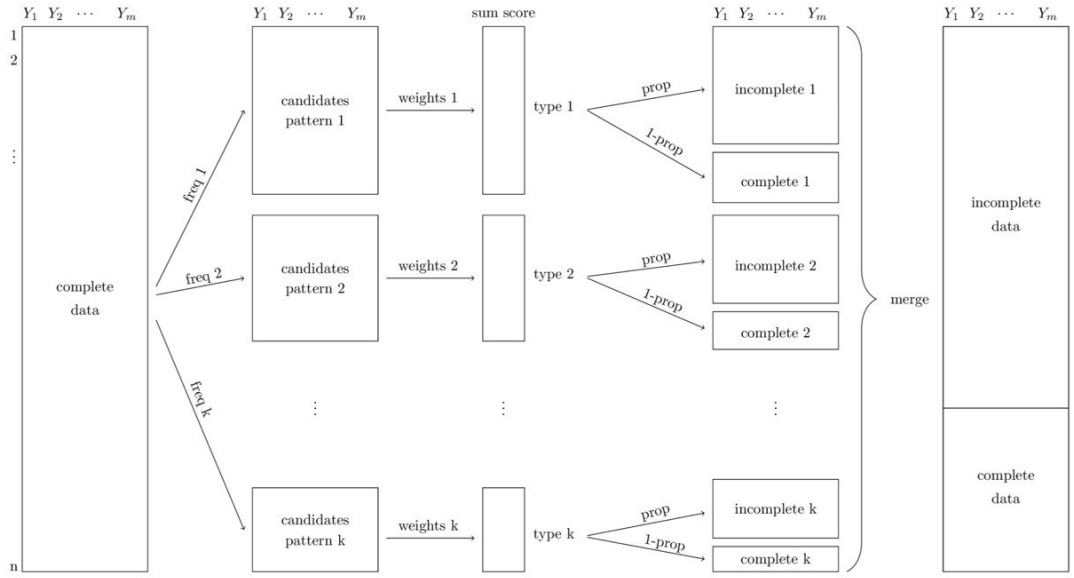


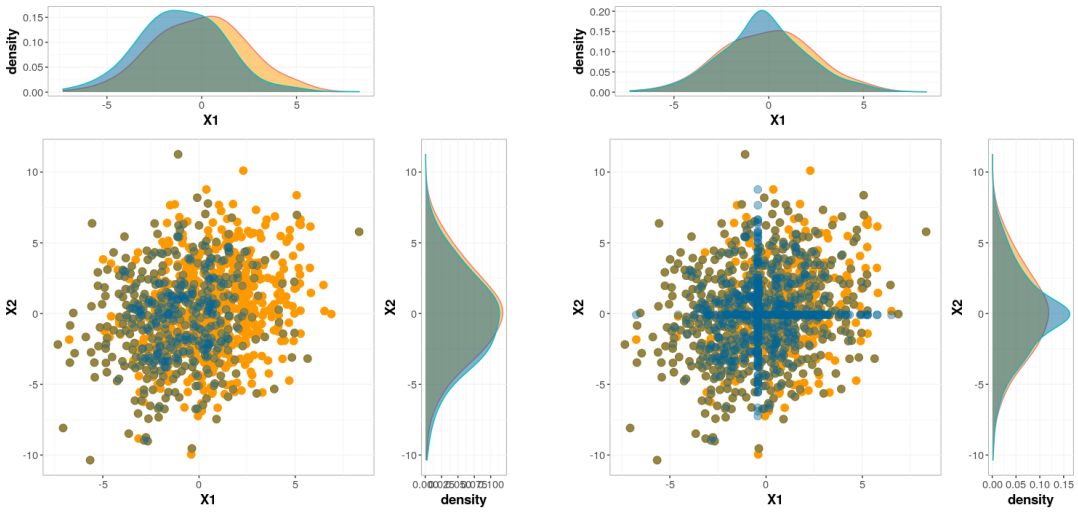
Рис. 5. Устройство метода ampute

Генерируем данные так же, как в случае с однофакторной потерей данных, симулируем пропущенные значение с помощью ampute. В итоге получим следующие значения корреляций (Рис.6):

cor_fulldate_X1X2	cor_MCAR_X1X2	cor_MAR_X1X2	cor_MNAR_X1X2
0.2469755	0.2476697	0.2522253	0.2544669

Рис. 6. Среднее значение корреляции между параметрами X_1 и X_2.

1.4 Простейшие методы работы с пропущенными данными



(a) Удаление пропущенных значений.

(b) Заполнение пропущенных значений средним.

Рис. 7

Последствиями простейших методов работы с пропущенными данными являются значительная потеря объема выборки, смещение оценок (Рис. 5), поэтому мы перейдем к рассмотрению множественного вменения.

2 Необходимые сведения из теории марковских процессов

Случайная последовательность — это такое отображение $X : \Omega \to \mathbb{R}^\infty$, что прообраз $\forall A \in \mathcal{B}(\mathbb{R}^\infty)$ лежит в сигма-алгебре $\mathcal{F}$, то есть $\{\omega : X(\omega) \in A\}$ является элементом $\mathcal{F}$.

Распределение случайной последовательности — это набор вероятностей $P(X \in B)$ для всех $B \in \mathcal{B}(\mathbb{R}^\infty)$.

Конечномерные распределения случайной последовательности — это набор вероятностей $p_{t_1,\dots,t_n}(B_1,\dots,B_n) = P(X_{t_1} \in B_1,\dots,X_{t_n} \in B_n)$ при $t_i \in \mathbb{N}, B_i \in \mathcal{B}(\mathbb{R}^\infty)$. В случае последовательностей целочисленных случайных величин задаём их распределение с помощью набора вероятностей $p_{i_1,\dots,i_n} = P(X_1 = i_1,\dots,X_n = i_n)$.

2.1 Сведения из теории марковских цепей

Случайная последовательность X_n, принимающая значения из конечного или счётного множества S, называется *марковской цепью* [6], если выполняется *марковское свойство*:

$$P(X_n = i_n | X_{n-1} = i_{n-1},\dots,X_0 = i_0) = P(X_n = i_n | X_{n-1} = i_{n-1}) = p_{i_{n-1}i_n}, \tag{8}$$

$\forall\, n = 0,1,\dots$ и $\forall\, i_0,\dots,i_n \in S$ с $P(X_{n-1} = i_{n-1},\dots,X_0 = i_0) > 0$, то есть любое испытание зависит от предшествующего и только от него.

Если вероятность перехода марковской цепи не зависит от n, то цепь *однородна по времени*, то есть все переходные вероятности не меняются со временем.

Говорят,что из состояния i *следует* состояние j, если $\exists\, n$:

$$P(X_n = j | X_0 = i) > 0, \tag{9}$$

то есть находясь в состоянии i возможно оказаться в j.

Матрица $\mathbf{P}$ называется *стохастической*, если $\forall i,j\ 0 \leqslant p_{ij} \leqslant 1$ и $\sum_j p_{ij} = 1\ \forall i$.
Распределение марковской цепи определяют:

- переходная стохастическая матрица цепи $\mathbf{P} = (P(X_t = i | X_{t-1} = j)) = (p_{ji})$,
- первоначальное распределение, т.е. распределение вероятностей в момент времени $t = 0$:

$$\mu = (\mu_1,\dots,\mu_n) = (P(X_0 = i_0)) = (P(X_0 = 1),\dots,P(X_0 = N)), \tag{10}$$

так что $\forall\, n = 0,1,\dots$ и $\forall\, i_0,\dots,i_n \in S$:

$$P(X_0 = i_0,\dots,X_n = i_n) = \mu_{i_0} p_{i_0 i_1},\dots,p_{i_{n-1}i_n}. \tag{11}$$

Увидим закономерность в преобразовании распределений. Рассмотрим распределение на первом шаге:

$$\begin{gathered} p^{(1)} = (p_1^{(1)},\dots,p_n^{(1)}) = (\mu_1 p_{1,1} + \dots + \mu_n p_{n,1}, \mu_1 p_{1,2} + \dots) = \\ (\mu_1,\dots,\mu_n) \begin{pmatrix} p_{1,1} & p_{1,2} & \dots \\ p_{2,1} & \ddots & \\ \vdots & & \end{pmatrix} = \mu\mathbf{P}. \end{gathered} \tag{12}$$

Таким образом, распределение на t-ом шаге:

$$p^{(t)} = (p_1^{(t)},\dots,p_n^{(t)}) = (P(X_t = i_t)) = \mu\mathbf{P}^t. \tag{13}$$

Цепь называется *неразложимой (неприводимой)*, если $\forall i, j$ из i следует j, а из j следует i. Иначе говоря, любые состояния $i_a, i_b \in S$ сообщаются, т.е. $\exists\ n$:

$$P(X_{m+n} = i_a | X_m = i_b) > 0. \tag{14}$$

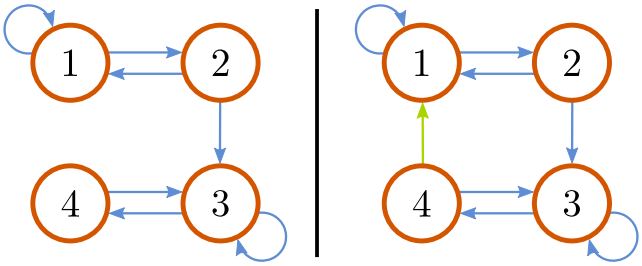


Рис. 8. Примеры разложимой (слева) и неразложимой (справа) цепей.

Неразложимая цепь называется *непериодической*, если наибольший общий делитель длин замкнутых путей в этой цепи равен 1. Иначе говоря, есть несколько путей из каких-либо состояний в себя, длины которых взаимно просты.

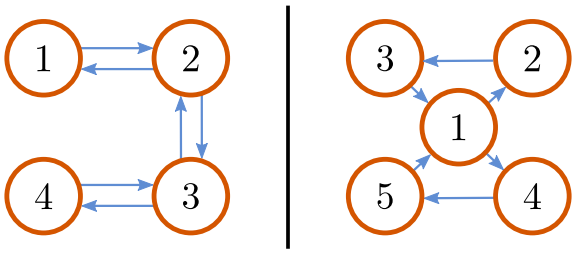


Рис. 9. Примеры периодичных цепей: слева с периодом k = 2, справа с k = 3.

Сведения, изложенные в пункте 2.1 будут использованы в следующем разделе.

2.2 Эргодическая теорема

Рассмотрим однородную марковскую цепь с двумя состояниями 0 и 1, которая задается матрицей переходных вероятностей

$$\mathbf{P} = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix}. \tag{15}$$

Если в выражении (6) положить $t = n$, то по индукции получаем

$$\mathbf{P}^n = \frac{1}{2 - p_{11} - p_{22}} \begin{pmatrix} 1 - p_{22} & 1 - p_{11} \\ 1 - p_{22} & 1 - p_{11} \end{pmatrix} + \frac{(p_{11} + p_{22} - 1)^n}{2 - p_{11} - p_{22}} \begin{pmatrix} 1 - p_{11} & -(1 - p_{11}) \\ -(1 - p_{22}) & 1 - p_{11} \end{pmatrix}. \tag{16}$$

В выражении (9) устремим $n \to \infty$, тогда

$$\mathbf{P}^n \to \frac{1}{2 - p_{11} - p_{22}} \begin{pmatrix} 1 - p_{22} & 1 - p_{11} \\ 1 - p_{22} & 1 - p_{11} \end{pmatrix}. \tag{17}$$

Таким образом, с течением времени $p_{ji}^{(n)}$ сходятся к предельным значениям π_i, не зависящим от j [10].

Далее приведем теорему из работы [10]:

Теорема 1. *Если конечная цепь Маркова с переходной матрицей* $\mathbf{P}$ *является неразложимой и апериодичной, то при любом начальном распределении* λ *вероятности* $P(X_n = i)$ *сходятся к некоторым* π_i, $i = 1, \ldots, N$ *при* $n \to \infty$*:*

$$\mathbf{P}^n \to \ \mathbf{\Pi}, n \to \infty. \tag{18}$$

При этом вектор $\pi = (\pi_1, \ldots, \pi_N)$ *является единственным решением системы уравнений*

$$\pi \mathbf{P} = \pi \tag{19}$$

и называется эргодическим (стационарным, инвариантным) распределением.

Таким образом, для неприводимой апериодической цепи Маркова с течением времени получаем распределение, не зависящее от первоначального:

$$\lim_{n \to \infty} P(X_n = j) = \pi_i \ \forall \lambda. \tag{20}$$

Следующая теорема с доказательством приведена из [7].

Теорема 2. *Пусть* X_n *— цепь Маркова. Следующие свойства эквивалентны:*

1. *Цепь обратима, т.е.* $\forall n \geqslant 1$ *и* $\forall$ *состояний* $i_0, \ldots, i_n$

$$P(X_0 = i_0, \ldots, X_n = i_n) = P(X_0 = i_n, \ldots, X_n = i_0); \tag{21}$$

2. *Цепь маркова* X_n *находится в состоянии равновесия* $(X_n \sim (\pi, \mathbf{P}))$*, где* π *— стационарное распределение для* $\mathbf{P}$*, или, другими словами,*
выполнены уранения детального баланса (сбалансированности) $\forall$ *состояний* i, j*, если:*

$$\pi_i p_{ij} = \pi_j p_{ji}. \tag{22}$$

Рис. 10. Два состояния обратимой марковской цепи.

Доказательство. Действительно, рассмотрим цепь с матрицей перехода $\mathbf{P}$ и начальным распределением λ:

1) $\Rightarrow$ 2). Пусть $n = 1$:

$$P(X_0 = i, X_1 = j) = P(X_0 = j, X_1 = i). \tag{23}$$

Просуммируем по j:

$$\sum_j P(X_0 = i, X_1 = j) = P(X_0 = i) = \lambda_i, \tag{24}$$

$$\sum_j P(X_0 = j, X_1 = i) = P(X_1 = i) = (\lambda\mathbf{P})_i. \tag{25}$$

Таким образом, $\forall i\ \lambda_i = (\lambda\mathbf{P})_i$, т.е. $\lambda = (\lambda\mathbf{P}) \Rightarrow$ цепь находится в состоянии равновесия со стационарным распределением λ. Далее,

$$\begin{gathered} P(X_0 = i, X_1 = j) = P(X_1 = j|X_0 = i)P(X_0 = i) = p_{ji}\lambda_i = \\ P(X_0 = j, X_1 = i) = P(X_1 = i|X_0 = j)P(X_0 = j) = p_{ij}\lambda_j, \end{gathered} \tag{26}$$

т.е. имеет место условие сбалансированности.

2) $\Rightarrow$ 1). Перепишем

$$P(X_0 = i_0, \ldots, X_n = i_n) = \pi_{i_0} p_{i_0 i_1} \ldots p_{i_{n-1} i_n} \tag{27}$$

и воспользуемся уравнением (11):

$$\begin{gathered} \pi_{i_0} p_{i_0 i_1} \ldots p_{i_{n-1} i_n} = \pi_{i_1} p_{i_1 i_0} \ldots p_{i_{n-1} i_n} = p_{i_1 i_0} \pi_{i_1} p_{i_1 i_2} \ldots p_{i_{n-1} i_n} = \\ p_{i_1 i_0} \pi_{i_2} p_{i_2 i_1} \ldots p_{i_{n-1} i_n} = \cdots = p_{i_1 i_0} p_{i_2 i_1} \ldots \pi_{i_n} p_{i_n i_{n-1}} = \pi_{i_n} p_{i_n i_{n-1}} \ldots p_{i_1 i_0} = \\ P(X_0 = i_n, \ldots, X_n = i_0) \blacktriangleright \end{gathered} \tag{28}$$

Условие сбалансированности является сильным инструментом для нахождения стационарного распределения:

Теорема 3. *Если λ и $\mathbf{P}$ удовлетворяют уровнениям детального баланса*

$$\lambda_i p_{ij} = \lambda_j p_{ji}, \tag{29}$$

то λ является стационарным распределением для $\mathbf{P}$, т.е. $\lambda\mathbf{P} = \lambda$.

Доказательство. Просуммируем по j:

$$\sum_j \lambda_i p_{ij} = \lambda_i \sum_j p_{ij} = \lambda_i, \tag{30}$$

$$\sum_j \lambda_j p_{ji} = (\lambda\mathbf{P})_i. \tag{31}$$

$\forall i$ выражения равны $\Rightarrow \lambda$ — стационарное распределение и цепь обратима по теореме 2.$\blacktriangleright$

Таким образом, если для заданного распределения p удастся найти такую матрицу перехода $\mathbf{P}$, что будет выполнено условие сбалансированности, то это распределение будет сходиться к стационарному с течением времени.

2.3 Марковские цепи с непрерывным множеством состояний

Основные идеи Марковской цепи с дискретным пространством состояний могут быть обобщены на непрерывный случай.

Марковским процессом X_n с дискретным временем и множеством (уже не обязательно не более чем счетным) состояний S будем называть такую последовательность, что

$$P(X_n \in A | X_{n-1} = x_{n-1}, \dots, X_0 = x_0) = P(X_n \in A | X_{n-1} = x_{n-1}). \tag{32}$$

Здесь $P(X_n \in A | X_{n-1}) = E(I_{X_n \in A} | X_{n-1} = x_{n-1})$ — условное математическое ожидание.

Функция $P(X_n \in A | X_{n-1} = x_{n-1}) = p(x_{n-1}, A, n)$ называется переходной функцией марковского процесса, если $S \subset \mathbb{R}$. Если существует плотность $f_{X_n|X_{n-1}}(y|x)$, то она называется переходной плотностью $P_{x,y,n}$.

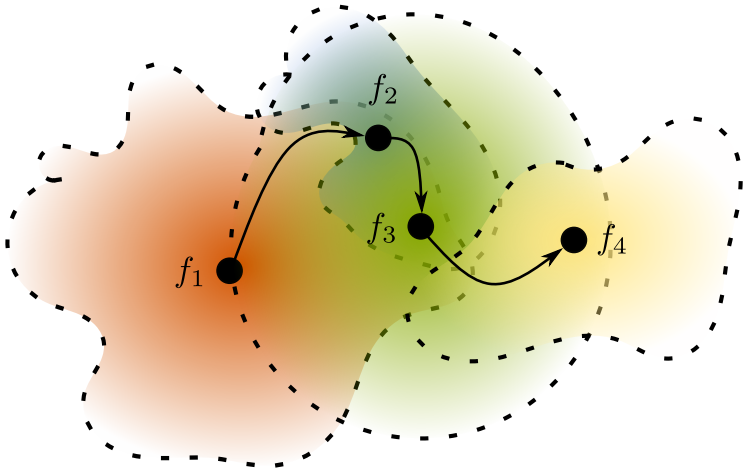


Рис. 11. Визуализация марковского процесса: новое состояние зависит от плотности распределения ему предшествующего.

Рассмотрим стационарное распределение π с переходной плотностью P, теперь уравнение Колмогорова - Чепмена [6] примет вид:

$$\int_x \pi(x) P(x, y) dx = \pi(y). \tag{33}$$

Обратимой назовем цепь, для которой выполняется уравнение детального баланса:

$$P(x, y) p(x) = P(y, x) p(y). \tag{34}$$

Опять условие сбалансированности влечет стационарность распределения.

Из эргодической теоремы для случая с непрерывным множеством состояний [11] мы получаем, что распределение процесса сходится к стационарному.

Тем самым, метод Монте-Карло будет работать и в случае марковских процессов с континуальным множеством состояний.

Общая идея метода Монте-Карло Марковских цепей (MCMC):

- построить цепь Маркова с эргодическим распределением, для которой стационарное распределение - это в точности заданное нами распределение,
- инициировать блуждание по цепи Маркова из некоторого начального состояния и дождаться, когда распределение сойдется к стационарному. С этого момента состояния цепи Маркова можно считать выборкой из желаемого распределения.

3 Гиббсовское семплирование

3.1 Идея и математическое обоснование

В некоторых типах задач одномерные условные распределения гораздо легче моделировать, чем совместные распределения. Идея гиббсовского семплирования заключается в том, что для восстановления совместного распределения рассматриваются только условные распределения для каждой переменной.

Для наглядности рассмотрим двумерную случайную величину (x, y). Мы хотим вычислить маргинальные распределения $p(x)$ и $p(y)$. Проще рассмотреть условные распределения $p(x|y)$, $p(y|x)$, чем искать их совместную плотность : $p(x) = \int p(x, y)dy$.

Семплирование начинается с некоторого значения y_0 для y. Генерируя случайную величину из условного распределения $p(x|y = y_0)$ получаем x_0. Далее используем x_0 для генерации нового значения y_1 из условного распределения $p(y|x = x_0)$. И так далее:

$$x_i \sim p(x|y = y_{i-1}),\ y_i \sim p(y|x = x_i). \tag{35}$$

Повторяя этот процесс k раз, получаем последовательность Гиббса длиной k с элементами (x_j, y_j), $1 \leqslant j \leqslant k$. Последовательность Гиббса сходится к стационарному распределению, которое и является искомым.

Рассмотрим случай нескольких переменных: пусть явный вид распределения вычислить трудно, но известны условные плотности

$$f_{X_i|X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_n}(x_i|x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n). \tag{36}$$

Нам нужно сгенерировать вектор $(X_1, \ldots, X_n)$, где мы знаем условные плотности.

- Выберем некоторый начальный вектор $(X_{1,1}, \ldots, X_{1,n})$, который вообще может получиться из нашего распределения.
- Выберем случайный индекс d из множества $1, \ldots, n$.
- Положим $X_{2,i} = X_{1,i}$ при $i \neq d$. Величину $X_{2,d}$ сделаем случайной с плотностью $f_{X_d|X_1,\ldots,X_{d-1},X_{d+1},\ldots,X_n}(x_d|x_1, \ldots, x_{d-1}, x_{d+1}, \ldots, x_n)$.
- Повторим шаг 2,3 для нового вектора и так далее.

Вектор $(X_{m,1}, \ldots, X_{m,n})$ сходится по распределению к требуемому вектору $(X_1, \ldots, X_n)$ при $m \to \infty$.

Алгоритм на каждом шаге берет случайную величину и задает ее значение при фиксированных остальных.

Таким образом, последовательно моделируются n случайных величин из n одномерных условных выражений вместо того, чтобы генерировать один n-мерный вектор за один подход с использованием полного совместного распределения. Последовательность генерируемых значений образует обратимую цепь маркова, эргодическое распределение которой является искомым.

Убедимся, что алгоритм гиббсовского семплирования укладывается в изложенную теорию. В данном случае мы рассматриваем процесс с переходной плотностью из $\vec{x} = (x_1, \ldots, x_n)$ в $\vec{y} = (x_1, \ldots, x_{k-1}, z, x_{k+1}, \ldots, x_n)$.

$$P(\vec{x}, \vec{y}) = \frac{1}{n} f_{X_k|X_1,\ldots,X_{k-1},X_{k+1},\ldots,X_n}(z|x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_n). \tag{37}$$

Проверим, что она удовлетворяет условию сбалансированности с нужной нам $p(\vec{x}) = f_{X_1,\ldots,X_n}(\vec{x})$. Действительно,

$$P(\vec{x}, \vec{y})p(\vec{x}) = \frac{1}{n} \frac{f_{X_1\ldots,X_n}(x_1 \ldots, x_{k-1}, z, x_{k+1}, \ldots, x_n)}{f_{X_1\ldots X_{k-1},X_{k+1},\ldots,X_n}(x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_n)} f_{X_1,\ldots,X_n}(x_1, ., x_n). \tag{38}$$

При этом $P(\vec{y}, \vec{x})p(\vec{y})$ даст ту же формулу с точностью до перестановки множителей. Следовательно, построенная цепь является сбалансированной с нужной плотностью.

3.2 Задача про одинаково распределенные точки в круге

Предположим, что мы хотим взять наугад выборку из N точек в единичном круге, удаленных друг от друга не менее чем на расстояние d. Выписать для данной задачи совместную плотность вектора трудно, однако условные плотности устроены довольно просто. Если мы знаем все точки, кроме одной, то оставшаяся точка распределена равномерно на исходном круге за вычетом кругов радиуса d вокруг остальных точек.

Следуя алгоритму гиббсовского семплирования :

- На первом шаге мы случайным образом выбираем точку в исходном круге радиуса 1 U_1. Назовем эту точку X_1. Пусть $U_d(X_1)$ — круг радиуса d с центром в точке X_1. Вторая точка выбирается случайно из множества $U_1 \setminus U_d(X_1)$. Третья точка X_3 выбирается случайно из множества $U_1 \setminus U_d(X_1) \setminus U_d(X_2)$. И так далее.
- Дальше мы генерируем случайный индекс i и меняем элемент X_i нашей выборки $(X_1, \ldots, X_N)$ на точку, равномерно распределённую в единичном круге без остальных $N-1$ кругов радиусов d.

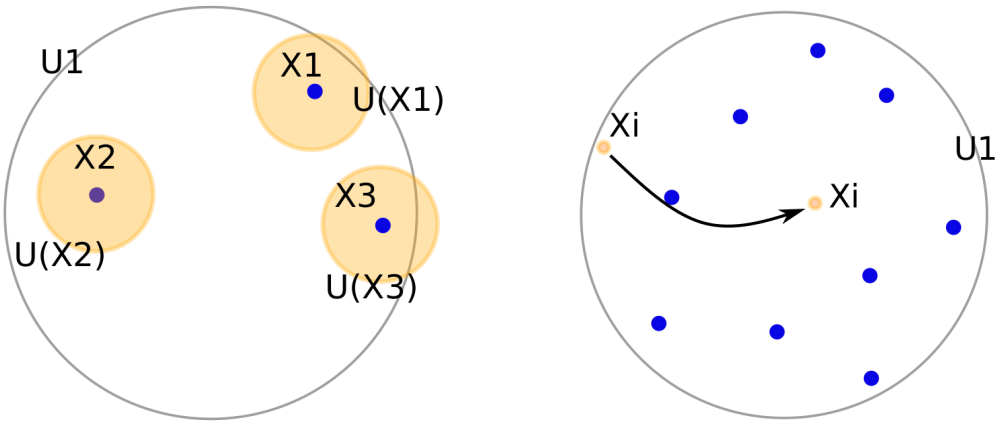


Рис. 12

После большого числа итераций полученный набор из N точек будет иметь искомое распределение:

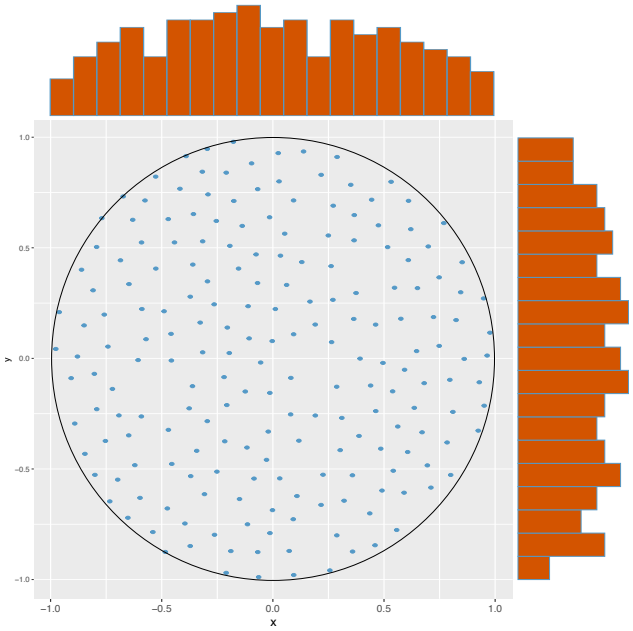


Рис. 13. Равномерное распределение точек в круге.

Если мы не будем использовать гиббсовское семплирование, то уже вторая сгенерированная точка не будет попадать в центр круга, поскольку её будет вытеснять равномерно - распределённая первая точка. Значит, первая и вторая точки не будут одинаково распределены.

Чтобы это заметить, пронаблюдаем распределение первой и второй брошенных в круг точек при гиббсовском семплировании и без (рис. 12 и рис. 13).

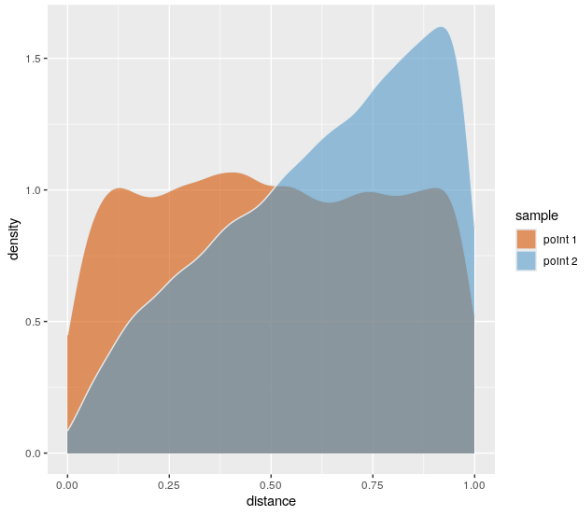


Рис. 14. Распределение точек без семпирования.

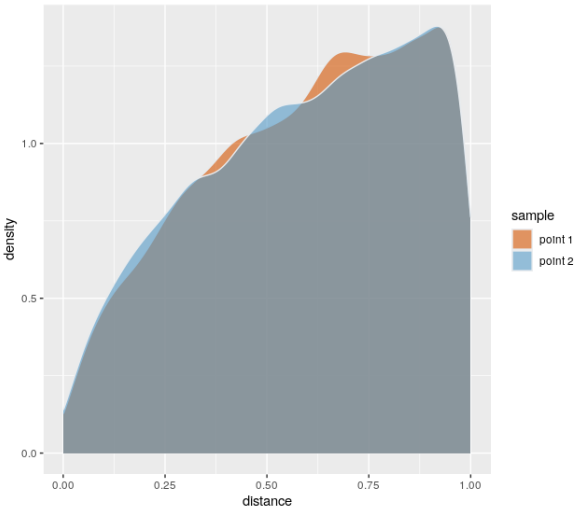


Рис. 15. Распределение с семплированием.

В рассмотренной задаче заполнения круга одинаково распределенными точками использование гиббсовского семплирования позволяет восстановить искомое распределение, исходя из условных плотностей. Задача восстановления данных имеет ту же структуру: совместное распределение посчитать сложно, но условные распределения приблизительно известны. Таким образом, изученный алгоритм может быть использован для решения задачи восстановления данных.

4 Множественное вменение методом цепных уравнений (MICE)

4.1 Алгоритм MICE

Многие из первоначально разработанных процедур множественного вменения предполагали совместное распределение для всех переменных. При больших наборах данных совместное распределение редко бывает уместным. MICE — это альтернативный подход к этим совместным распределениям, поскольку алгоритм в MICE — это алгоритм семплирования Гиббса. Он начинается со случайной выборки из наблюдаемых данных и циклически перебирает условные распределения.

Пусть $Y_1, Y_2, ..Y_p$ — переменные с пропущенными значениями.

- 0. Для всех Y_j:
    1. Указать модель заполнения $P(Y_j^{mis}|Y_j^{obs}, Y_{-j})$;
    2. Выбрать первоначальное заполнение Y_j^0 случайным образом из Y_j^{obs};
- Итерации $t = 1, \ldots, M$:
    1. Определяем $\dot{Y}_{-1}^t = (\dot{Y}_2^{t-1}, \ldots, \dot{Y}_p^{t-1})$;
    2. $\dot{\theta}_1^t \sim P(\theta_1^t|Y_1^{obs}, \dot{Y}_{-1}^t)$;
    3. $\dot{Y}_1^t \sim P(Y_1^{mis}|Y_1^{obs}, \dot{Y}_{-1}^t, \dot{\theta}_1^t)$;
    4. ...
    5. $\dot{Y}_{-j}^t = (\dot{Y}_1^t, \ldots, \dot{Y}_{j-1}^t, \dot{Y}_{j+1}^{t-1}, \ldots, \dot{Y}_p^{t-1})$;
    6. $\dot{\theta}_j^t \sim P(\theta_j^t|Y_j^{obs}, \dot{Y}_{-j}^t)$;
    7. $\dot{Y}_j^t \sim P(Y_j^{mis}|Y_j^{obs}, \dot{Y}_{-1}^t, \dot{\theta}_j^t)$;
    8. ...
    9. $t++$;

Количество итераций перебора задаётся пользователем, либо итерации происходят до тех пор, пока, наконец, не получат совместное распределение переменных.

Ван Бюрен установил достаточное количество итераций в пределах от 5 до 20.

4.2 Множественное вменение

Множественное вменение — это общий подход к проблеме пропущенных данных, направленный на то, чтобы учесть неопределенность в отношении недостающих значений путем создания нескольких различных наборов данных.

Есть некоторая θ (scientific estimand) — оценка, которую мы сможем вычислить только в том случае, если нам известны данные о всей популяции, чего почти никогда не бывает. Цель множественного вменения — найти достоверную оценку этой θ. Для понимания, это может быть среднее значение популяции, корреляции, коэффициенты регрессии.

Множественное вменение состоит из 3х этапов: заполнения, анализа, объединения.

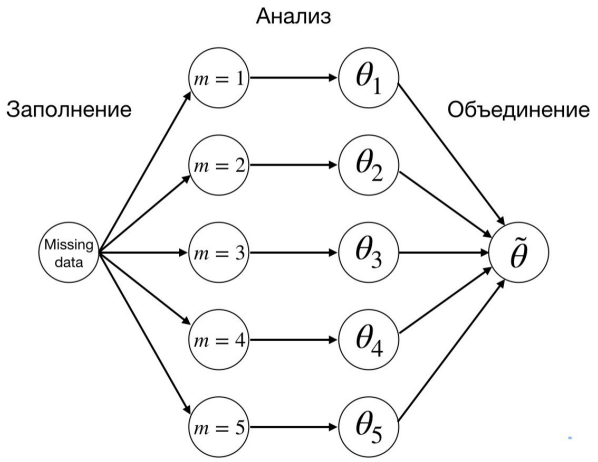


Рис. 16

На первом шаге создаются m копий исходного датасета (в примере их 5), которые далее параллельно заполняются рассмотренным ранее алгоритмом гиббсовского семплирования.

Сколько копий датасета нужно параллельно заполнять и оценивать? В своей работе 1987 года [9] Рубин показал, что всего пять наборов данных приведут к эффективности более 90%.

После того, как 5 наборов данных заполнены, в них проводится анализ, то есть поиск необходимого параметра, после чего все оценки объединяются по правилам Рубина.

4.3 Результативность метода

Посмотрим на результаты сравнения методов complete case analysis, стохастического заполнения и метода MICE. Сравнение проводится таким образом:

Генерируем 1000 раз полный датасет с $\mu = (5, 5, 10)$, $\sigma = \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix}$ для $\rho \in (0.1, \ldots, 0.9)$.

Каждый раз симулируем в датасете пропущенные данные методом ampute, после чего применяем методы заполнения complete case analysis, стохастическое заполнение, и MICE и оцениваем смещение среднего относительно настоящих (полных) значений. После усреднения 1000 значений смещения для каждого ρ для каждого метода заполнения получаем:

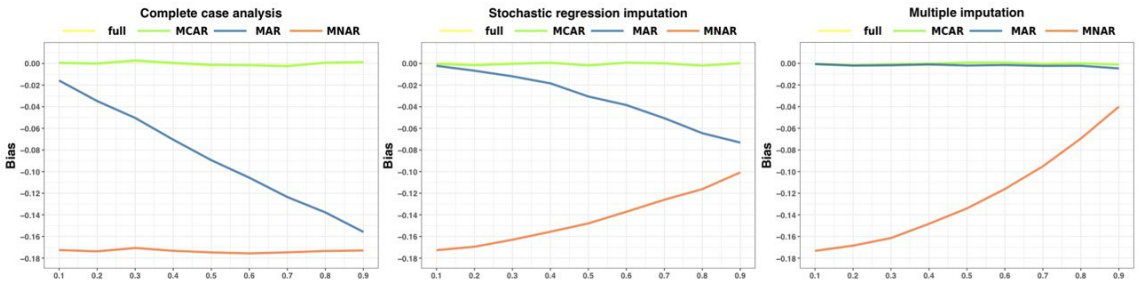


Рис. 17. Ось x — корреляции внутри исходных данных $\rho \in (0.1, \ldots, 0.9)$

Те же действия проводим для поиска доверительного интервала:

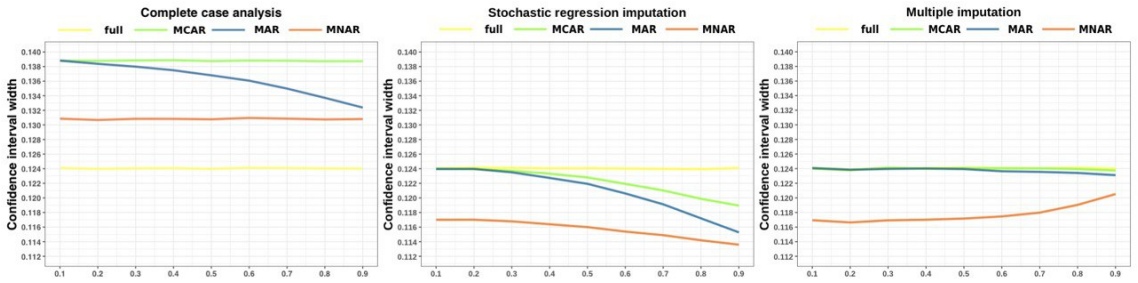


Рис. 18. Ось x — корреляции внутри исходных данных $\rho \in (0.1, \ldots, 0.9)$

По результатам видно, что при заполнении пропущенных данных с помощью MICE при типах потерянных данных MCAR, MAR смещение отсутствует и доверительный интервал совпадает с настоящим (исходным). Даже для такого сложноустроенного типа пропущенных данных, как MNAR, ситуация улучшается по сравнению с другими методами: с ростом ρ смещение уменьшается и значение ширины доверительного интервала близиться к оригинальному.

Вывод

Мы показали, что можно построить марковскую цепь с заранее заданным эргодическим распределением. В эту теорию укладывается метод гиббсовского семплирования, который позволяет моделировать совместное распределение, используя только условные, что облегчает вычисления во многих задачах. Например, в рассмотренной задаче заполнения круга одинаково распределенными точками использование гиббсовского семплирования позволяет восстановить искомое распределение, исходя из условных плотностей. Задача восстановления данных имеет ту же структуру: совместное распределение посчитать сложно, но условные распределения приблизительно известны.

В работе получены результаты, демонстрирующие, что проблема пропущенных данных может быть эффективно решена с помощью метода MICE, то есть путём создания нескольких вариантов заполнения данных с помощью алгоритма гиббсовского семплирования и анализа всех этих вариантов в одно наилучшее заполнение.

Для данной работы алгоритм гиббсовского семплирования и графики были реализованы с помощью языка программирования R. Код доступен по ссылке [12]. Иллюстрации выполнены в графическом редакторе Inkscape.

Список литературы

[1] Van Buuren S. Flexible imputation of missing data. – CRC press, 2018.

[2] Муромцева, Г. А., et al. Распространенность факторов риска неинфекционных заболеваний в российской популяции в 2012-2013гг. Результаты исследования ЭССЕ-РФ. Кардиоваскулярная терапия и профилактика 13(6) (2014).

[3] Rianne Margaretha Schouten, Peter Lugtig Gerko Vink (2018): Generating missing values for simulation purposes: a multivariate amputation procedure, Journal of Statistical Computation and Simulation

[4] Buuren, S. V., Groothuis-Oudshoorn, K. Mice: Multivariate imputation by chained equations in R. Journal of statistical software (2010).

[5] Metropolis, N., Ulam, S. The monte carlo method. Journal of the American statistical association, 44(247) (1949).

[6] Феллер, В. Введение в теорию вероятностей и ее приложения (Vol. 2). Рипол Классик (2013).

[7] Кельберт, М., Сухов, Ю. Вероятность и статистика в примерах и задачах. Том 2. Марковские цепи как отправная точка теории случайных процессов и их приложения. Litres (2017).

[8] Murphy, K. P. Machine learning: a probabilistic perspective. MIT press (2012).

[9] Rubin D. Multiple imputation for nonresponse in surveys. New York: John Wiley and Sons; 1987.

[10] Ширяев, А. Н. Вероятность. В 2-х кн. Москва: МЦНМО (2004).

[11] Bellet, L. R. Ergodic properties of Markov processes. In Open quantum systems II (pp. 1-39). Springer, Berlin, Heidelberg (2006).

[12] `https://github.com/annalimm/Gibbs_circle.git`.